Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 8: 25–44 (2024)
DOI: 10.3897/mbmg.8.108649

Research Article

Simple approaches for evaluation of OTU quality based on dissimilarity arrays

Marie-Josée Cros[1], Jean-Marc Frigerio[2,3], Nathalie Peyrard[1], Alain Franc[2,3]

1 Université de Toulouse, INRAE, UR Mathématique et Informatique Appliquées Toulouse, 24 chemin de Borde Rouge, 31320 Auzeville-Tolosane, France
2 INRAE, Université de Bordeaux, BIOGECO, 64 route d'Arcachon, 33612 Cestas, France
3 INRIA BSO, Université de Bordeaux, 200 avenue de la vieille tour, 33405, Talence, France
Corresponding author: Nathalie Peyrard (nathalie.peyrard@inrae.fr)

OPEN ACCESS

Academic editor: Gert-Jan Jeunen
Received: 3 July 2023
Accepted: 14 February 2024
Published: 14 March 2024

Citation: Cros M-J, Frigerio J-M, Peyrard N, Franc A (2024) Simple approaches for evaluation of OTU quality based on dissimilarity arrays. Metabarcoding and Metagenomics 8: e108649. https://doi.org/10.3897/mbmg.8.108649

Abstract

An accurate and complete taxonomic description of the diversity present in an environmental sample is out of reach at this time. Instead, metabarcoding is used today and it is expected that OTUs represent a category relevant for biodiversity inventories on a molecular basis. However, artefacts in the production of OTUs can occur at different stages and may impact ecological conclusions. We propose to evaluate the quality of OTUs in a sample by characterising the deviation of each OTU's dissimilarity array from that of an ideal OTU where all sequences are at distances smaller than the barcoding gap. We consider two deviations: the creation of composed OTUs, corresponding to the artificial merging of several OTUs and the creation of noisy OTUs that contain some sequences that are loosely associated with the core sequence of the OTUs and that do not form a compact subgroup. We propose a simple and automatic 2-step method that successively categorises the OTUs of a sample as composed or single and then identifies OTUs with noise amongst the single ones. The associated code is available at https://forgemia.inra.fr/alain.franc/otu_shape. We applied the method on 32 samples of diatoms from Arcachon Bay (France) that represent contrasted environmental conditions and we obtained good agreement with expert categorisation of OTUs. We suggest that single OTUs without noise can be used as such for further ecological studies. Composed OTUs should be post-treated with classical clustering or community detection tools. The quality of single OTUs with noise remains to be further tested via supplementary studies on a diversity of organisms.

Key words: Composed OTU, diatoms, metabarcoding, OTU with noise, support vector machine, stochastic block model

Introduction

Exponential development of Next Generation Sequencing and High Throughput Sequencing has facilitated the mass production of barcodes in environmental samples with metabarcoding (Hajibabaei et al. 2011; Taberlet et al. 2012; Kermarrec et al. 2013), produced in bulk, without knowing which organism they come from, especially in microbial communities. An environmental sample in metabarcoding is a set of reads that are representative of the diversity of the community that has been sampled and an approach with on-going very promising developments for diversity studies. An Operational Taxonomic Unit (OTU) is a set of sequences that are ideally at a distance smaller than a given level referred to as barcoding gap (Blaxter et al. 2005). Being a set of sequences close to each other, it is expected that OTUs identified in an environmental sample represent a category relevant for biodiversity inventories on a molecular basis, where assemblages of OTUs mimic the organisation of communities as assemblages of species. Note that this raises the question of qualifying and quantifying a correspondence (or not) between OTUs and the notion of species, which has been the subject of a long debate. Keeping in line with Blaxter et al. (2005), we adopt here the view that we are “agnostic as to whether the taxa we can define using these barcode sequences [...] are species or not ”. In our work, an OTU is defined as a set of sequences that are mutually close and there is no attempt to make sense of an OTU, for example, by naming it.

OTUs are building blocks of molecular-based inventories and there are various protocols for building them from sets of sequences in an environmental sample. Artefacts in the production of OTUs can occur at different stages (see, for example, Bik et al. (2012)). Moreover, as OTUs are used downstream for computing diversity indices or performing statistical ecology, different delineations between OTUs may lead to different diversity indices or ecological profiles. The impact on diversity studies has been studied thoroughly and some tools already exist to clean OTUs. For instance, there may be more OTUs than expected from the expert knowledge about the diversity of the system studied. In Froslev et al. (2017), the authors propose a post-treatment method to identify and merge redundant OTUs, based on the identification of sequences similarities and of systematic co-occurrence of the OTUs in multiple samples. With metabaR (Zinger et al. 2021), it is possible to remove artefactual OTUs, based on the analysis of their abundance across different samples. On the contrary, sequences of two distinct OTUs can be artificially grouped. For instance, it is known that single linkage clustering leads to chaining effects that may lead to the merging of two or more OTUS. In SWARM (Mahé et al. 2014, 2015), a post-treatment is proposed, referred to as the breaking phase, to split the potentially composed OTU. This is done by exploring the inner structure of OTUs which, for a composed OTU, is formed by peaks of abundant amplicons with a valley in between (one peak per entity). We propose to complete these tools for post-treating the OTUs of a sample, by using only the array of pairwise dissimilarities between sequences in each OTU.

To characterise the notion of quality of an OTU, we refer to an ideal OTU (where all dissimilarities within an OTU are smaller than the barcoding gap) and we identify possible deviations from the theoretical pattern of the corresponding dissimilarities array. Deviations, when they exist, are not random. We study two deviations leading to composed OTUs and OTUs with noise. As defined above, composed OTUs are the artificial merging of several OTUs, as opposed to single OTUs. We propose a new way to identify composed OTUs. Unlike the breaking phase in Mahé et al. (2014, 2015), it does not rely on a threshold parameter that must be fixed arbitrarily. Then, once composed OTUs have been split into single OTUs, we consider a second post-treatment to identify the presence of noise. We say that an OTU contains noise if it contains some sequences that are loosely associated with the core sequences and that do not form a compact subgroup of sequences. To the best of our knowledge, the identification and quantification of noise in OTUs has seldom been addressed. Our approach is a classification method, based on simple statistics derived from the dissimilarity matrix and on learning methods like a linear Support Vector Machine (SVM, Cortes and Vapnik (1995)) and a Stochastic Block Model (SBM, Daudin et al. (2008)).

We apply the approach on a dataset of diatoms from Arcachon Bay, kindly made available by the Malabar project (Auby et al. 2022 and https://entrepot.recherche.data.gouv.fr/dataverse/malabar). We believe that the fraction of the three OTU types (composed, single without noise and single with noise) present in an environmental sample can provide knowledge about the ecology of the sample. As an illustration, we present results about the dependencies between the fraction of types and some known environmental variables describing the conditions under which the diatoms samples were collected.

Method

Data

The material required as input of our method for identifying OTU types of a sample is the list of OTUs in the sample together with the array *D* of pairwise dissimilarities associated with each OTU. Rigorously, the mathematical object is a matrix. However, here and in the following, we use the term array, by reference to the operational implementation of the building of *D*, since in the code, it is a variable of type array.

Samples

We apply our method on a dataset of diatoms from Arcachon Bay. They represent a sampling of the diversity of photosynthetic protists, mainly diatoms, in Arcachon Bay, France. Samples are allocated equally amongst the four seasons (autumn, winter, spring, summer), four locations (Bouée 13, Comprian, Jacquets, Teychan) and two water columns (pelagic high tide and benthic). This yields 4 x 4 x 2 = 32 samples. Sample sizes range between 19,000 and 36,000 reads (Suppl. material 1: D, table S1). Reads are amplicons of a 312 bp region in the rbcL marker (see Rimet et al. (2016)).

Dissimilarity array

For each sample, pairwise dissimilarities after dereplication between reads have been computed with the Smith-Waterman local alignment score. Due to the size and number of fasta files, we have used the distributed version of disseq called mpidisseq (see https://gitlab.inria.fr/biodiversiton/disseq), run on the cluster CURTA of the mésocentre of Nouvelle-Aquitaine. Hence, a *n* by *n* dissimilarity array is attached to each sample if it is composed of *n* reads.

List of OTUs

The dissimilarity array of a sample is denoted *D* and the dissimilarity between reads *i* and *j* is denoted *d*(*i*, *j*) (term at row *i* and column *j* of *D*). In a second step, we computed OTUs from *D* for each sample. The numbers of reads and OTUs per sample are given in Suppl. material 1: D, table S1. Two reads *i* and *j* belong to the same OTU if their dissimilarity is smaller than a selected barcoding gap (Blaxter et al. 2005). Therefore, we implemented the following procedure on the dissimilarity array *D* after dereplication:

1. Select a barcoding gap *g* (here, $g = 9$, representing 3% of the marker length);
2. Create a graph $G = (V, E)$, where nodes $i \in V$ are the reads in the sample and $(i, j) \in E$ if and only if $d(i, j) \leq g$;
3. Compute all connected components of *G*.

An OTU is then defined as a connected component of *G*. The associated subgraph of *G* is denoted G_{otu}. It is connected by construction, but it is not always a clique since we can have three elements i,j,k such that d(i,j) ≤ g and d(j,k) ≤ g (therefore i, j and k are in the same connected component), but d(i,k) > g (the barcoding gap). It is the well-known chaining effect. For each sample, we extracted one dissimilarity array per OTU, denoted D_{otu}. Hence, we worked with 32 sets of dissimilarity arrays. We kept the OTUs with 15 reads or more only, because it would not be meaningful to try to identify groups in smaller OTUs.

We checked that the OTUs obtained are very close to the outputs of SWARM. It is not surprising because our procedure relies on building connected components at a given threshold and this is known to be equivalent to hierarchical aggregative clustering with Single Linkage (Gower and Ross 1969). SWARM relies on a bottom-up algorithm (aggregate with seeds) equivalent to clustering with single linkage. When comparing our OTUs with SWARM, we noticed that each SWARM OTU was included entirely within one of our OTUs. The difference is due either to the breaking phase in SWARM which divides some of our OTUs or to the production by SWARM of a long tail with many very small sets of sequences, in particular many singletons.

Annotated reads

A reference database for the rbcL marker for diatoms is available (Rimet et al. 2016). We mapped each read of the whole sample, regardless of the OTU it belongs to, on this reference database, with the diagno-syst tool (Frigerio et al. 2016). This algorithm first calculates all the distances between the reads in the sample and the sequences in the reference database, retaining only the pairs below a certain threshold. It then lists the affiliations of the references retained for a given read and transfers the taxonomic affiliation from the reference database to the read if it is homogeneous in this list. If there are several different affiliations, the read is described as “ambiguous”. Therefore, we were able to explore the taxonomic profile of some of the OTUs typed as being composed or with noise. Not all reads reached a match. For this study, we have limited ourselves to OTUs with all reads annotated in order to have a complete knowledge of the species present in the OTUs. This was the case for 180 OTUs and the large majority (about 85%) were monospecific. However, it is worth noting that we have found one reference sequence of *Rhizosolenia fallax* which is present once and once only in several fully annotated OTUs belonging to several genera. We ignored the sequence and declared the OTU as monospecific. Note that the annotated reads are not data used by our method for typing OTUs. We only use this information for validation.

Ideal OTUs and deviations

Ideal OTU

A drawback of the above 3-step procedure for building the OTUs is that two reads can have a dissimilarity larger than *g* and still belong to the same OTU. Is there a way to define an OTU with the same procedure, but with the guarantee that all dissimilarities within an OTU are below the barcoding gap? In this case, G_{otu} is a clique. The answer is yes if we use a dissimilarity such that d(i,j) ≤ g and d(j,k) ≤ g implies that d(i,k) ≤ g (it means that the relationship defined by "*i relates to j* if and only if *d*(*i, j*) ≤ *g*" is transitive). This is possible if and only if *d* is a distance and is ultrametric. A distance *d* is said to be ultrametric if it fulfils the condition *d*(*i, j*) ≤ max(*d*(*i, k*), *d*(*j, k*)) for any read *k*, which is stronger than the classical triangular inequality. Dissimilarities computed as edit distances between two reads are not ultrametric and, therefore, the relationship defined as being at a distance below a barcoding gap is not transitive. On the contrary, the age of the Most Recent Common Ancestor (MRCA) between two reads is ultrametric. If *D* is built with the MRCA as distance and steps 1 to 3 above are applied, all connected components of *G* are cliques, and an OTU is a clique. Such an OTU is said to be "ideal".

Our hypothesis is that the observed deviations from this ideal OTU structure are not random, but are themselves structured. In what follows, using only the dissimilarity arrays, we describe two ways in which an OTU can diverge from being ideal: composed OTU and OTU with noise.

Composed OTU

First, we define what is a single OTU. A single OTU is close to what would be a theoretically ideal OTU, where all dissimilarities in D_{otu} are smaller than the barcoding gap. There may be a few exceptions for some sequences, but we will deal with that in a second step, when defining OTU with noise. The corresponding graph G_{otu} is composed of a single large strongly connected entity with the possibility of some satellite nodes. Composed OTUs deviate from ideal and single OTU by the fact that they correspond to dissimilarity arrays with a structure of two or more blocks, with intra-block dissimilarities smaller than the barcoding gap and most of the inter-block dissimilarities larger. This leads to a graph G_{otu} with several entities, where the nodes in an entity are strongly connected and there are few connections between the entities. In graph theory, such a graph is said to have a community structure (Girvan and Newman 2002) and each entity is a community. In Fig. 1, we provide an example of a graph of an ideal OTU, of a single OTU with satellite nodes and of a composed OTU.

It is well known that composed OTUs can be produced during the phase of clustering of the sample reads due to the above mentioned chaining effect. It usually corresponds to the grouping of reads from different species in the same OTU. We illustrate this chaining effect on a sample by comparing the species and the OTU that each sequence belongs to. In Fig. 2, we show twice the same point cloud: the projection on the first two axes by Multidimensional Scaling (MDS, Cox and Cox (2001)) of the dissimilarity array of a sample, once where dots (i.e. reads) are coloured according to the OTU they belong to and once according to the species they belong to. It is clear that the blue OTU (the largest) is composed because of the archipelago of isolated dots scattered amongst the three entities, which leads to chaining and a spurious OTU.

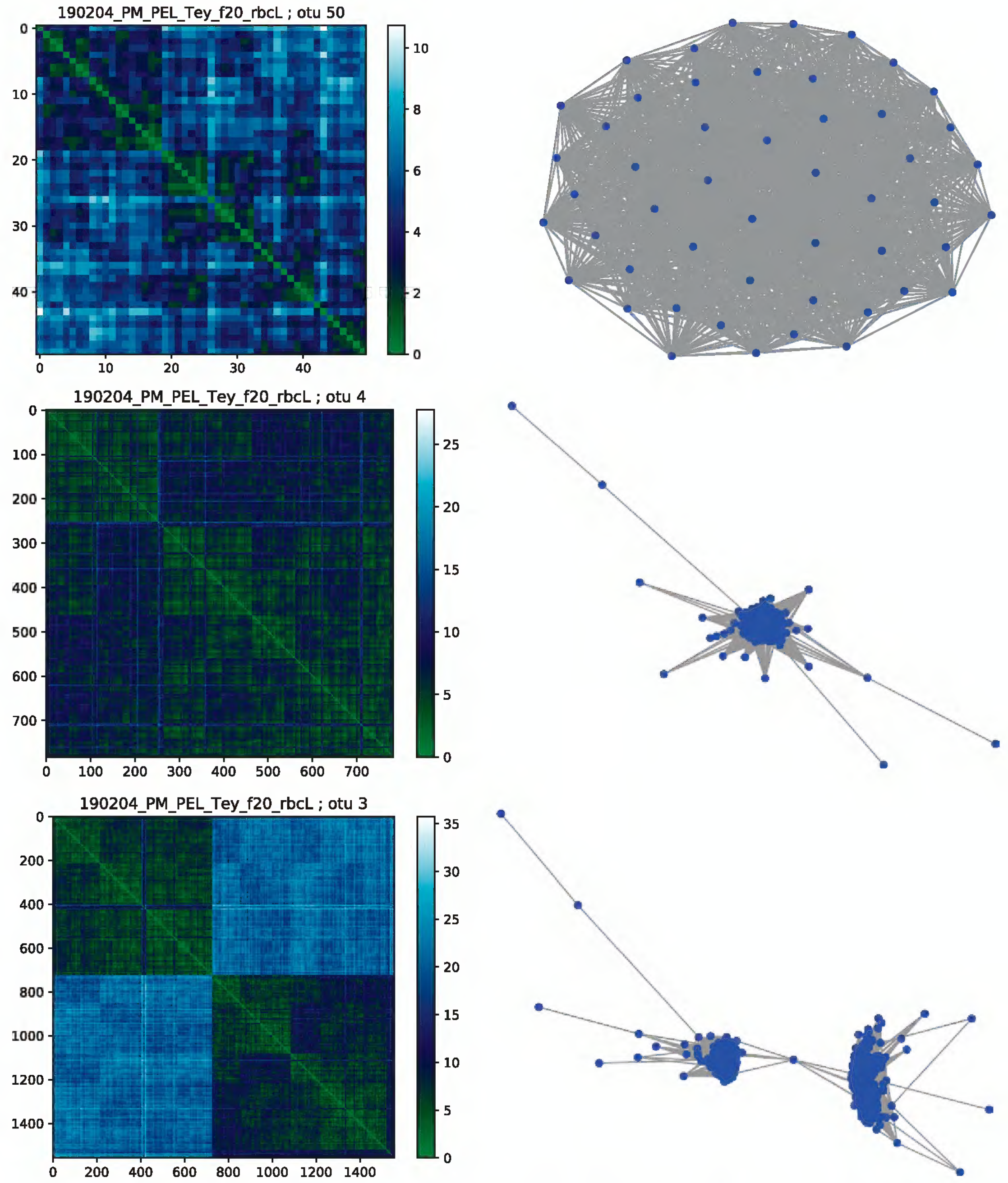


Figure 1. Examples of graph G_{otu} for three types of OTUs, from top to bottom: (i) ideal OTU, which is single and a clique (each read has a dissimilarity smaller than the barcoding gap with all the other reads of the OTU; (ii) a single OTU with a large strongly connected core entity and some satellite nodes; (iii) a composed OTU, consisting of several entities with high intra-entity connections rates and low between entities connection rates (and some satellite nodes as well).

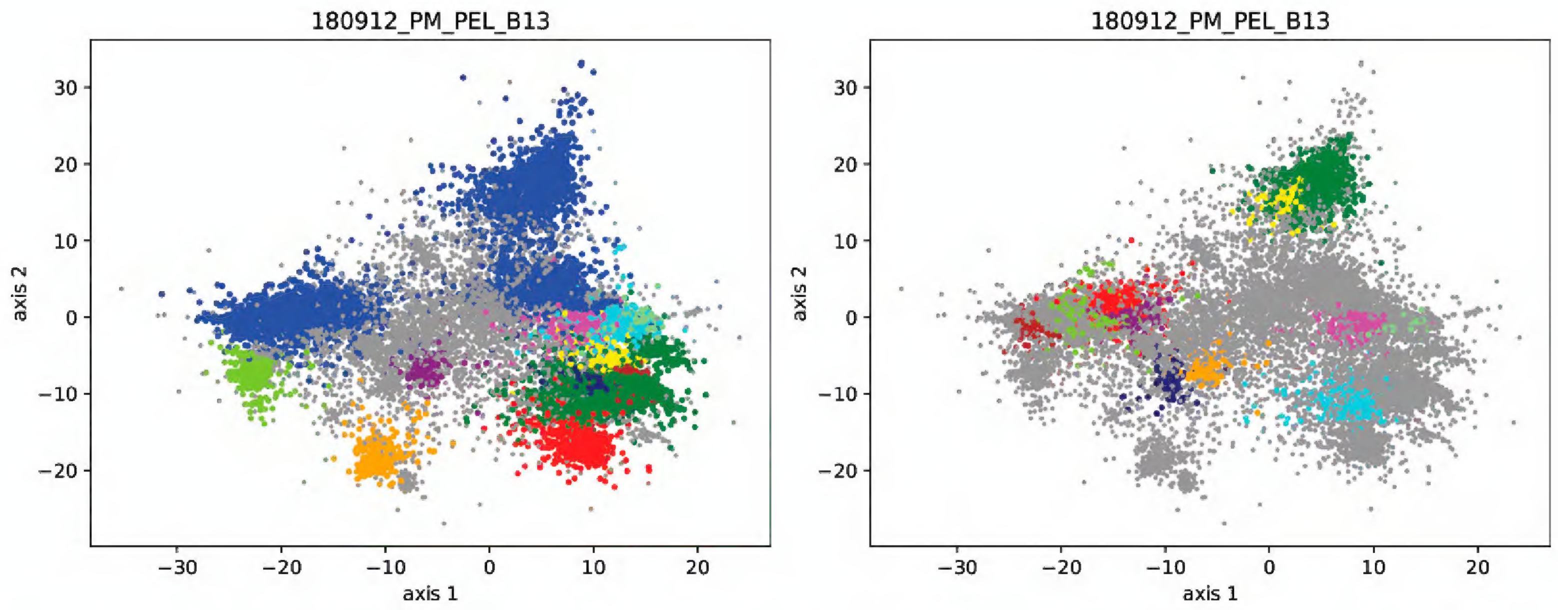


Figure 2. Illustration of the chaining effect. Both figures display the same scatter plot of sample 180912_PM_PEL_B13 (high tide, pelagic, summer, Bouée 13), where one dot is a read (there are 37036 dots), with the first MDS component on the x axis and the second one on the y axis. The two plots differ by the way dots are coloured. In the left plot, dots are coloured according to the OTU they belong to. In the right plot, they are coloured according to the species they have been assigned to. Only the species and OTUs with the 12 largest sizes have been coloured; the remaining ones are coloured in grey (if not, many colours would have been indistinguishable).

OTU with noise

Then, amongst single OTUs, we describe a second deviation from an ideal OTU: OTU with noise. For these, there could still be some reads that are loosely associated only with the rest of the OTU and that are too far from each other to form themselves an entity: for such a sequence i, dissimilarities $d(i, j)$ are below the barcoding gap for only a small number of sequences j. These sequences are far from the core sequences of the OTU and they do not form a second entity (as in a composed OTU) since they can be far from each other (see Fig. 1, centre). We qualify these sequences as composing the noise.

In the following, we present a simple and automatic 2-step method that successively categorises the OTUs of a sample as composed or single and then identifies OTUs with noise amongst the simple ones. We will show that composed and single with noise OTUs represent the majority of the OTUs in the different samples of our dataset.

Method for identifying composed OTUs

In a first step, we propose an automatic unsupervised method for sorting the OTUs of a sample into two groups: single ones and composed ones. In a single OTU, most dissimilarities in D_{otu} will be smaller than the barcoding gap. For a composed OTU, there will be a significant proportion of dissimilarities larger than the gap (due to the inter-entity dissimilarities). This is the information we use to discriminate between single and composed OTUs. For a given OTU, we build G_{otu} from D_{otu}. We then define θ as the ratio between the number of missing edges in G_{otu} and the total number of possible edges. The number of missing edges corresponds to half the number of elements in D_{otu} that are larger than the barcoding gap (since D_{otu} is a symmetric array and each dissimilarity appears twice). It is equal to $\Sigma_{i<j}\ \delta\ (d_{otu}(i, j) > g)$ where the sum is over all pairs (i, j) of lines and columns of D_{otu} where $i < j$. The function δ is equal to 1 if the condition is satisfied and 0 otherwise. The total number of possible edges in G_{otu} is equal to $\frac{(n_{otu} (n_{otu} - 1)}{2}$, where n_{otu} is the number of reads in the OTU. Therefore, $\theta = 2\ \Sigma_{i<j}\ \delta\ (d_{otu}(i, j) > g)/(n_{otu}\ (n_{otu} - 1))$. Then, for single OTUs, θ will be small, because very few edges are missing. For composed OTUs, θ will be large. Indeed, let us take as an example an OTU with two balanced entities. There will be few missing edges within each entity, but many edges missing between both entities. If each entity has n_{otu} /2 sequences, there are possibly n^2_{otu} /4 edges between both entities and as many potential missing edges. Hence $\Sigma_{i<j}\ \delta\ (d_{otu}(i, j) > g) \approx n^2_{otu}$ /4 while $n_{otu}\ (n_{otu} - 1) \approx n^2_{otu}$. Finally, $\theta \approx 1/2$.

To sort the OTUs of a sample into composed and single ones, we use θ, which can be computed directly from D. We define a critical value θ_c as follows. We compute θ for each OTU and we build a smoothed version of the histogram of the θs using a Gaussian kernel (see Suppl. material 1: B). This estimated density always shows a first large mode around low values of θ, followed by one or several other less important modes. We define θ_c as the value of θ for which the minimum of the estimated density is reached between the first and the second mode. If $\theta < \theta_c$, the OTU is classified as single, otherwise it is classified as composed (see Fig. 3).

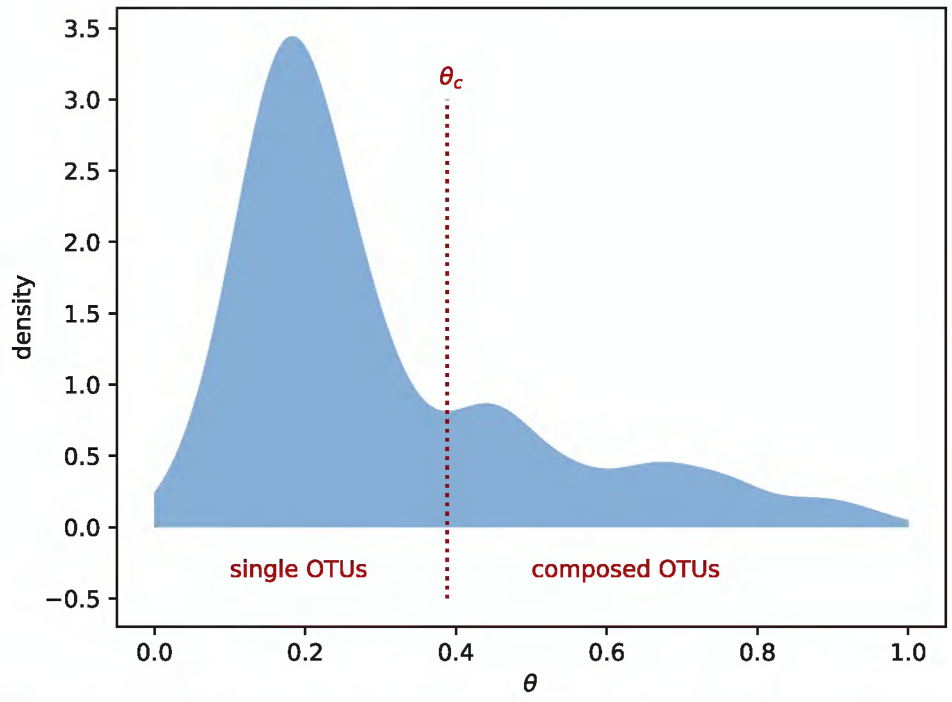


Figure 3. Principle of the method for sorting OTUs of a sample into composed and single ones. Example of a smoothed version of the histogram of θ values (ratio between the number of missing edges in G_{otu} over the total number of possible edges in the OTU): θ_c is the first local minimum after the first mode, OTUs with $\theta < \theta_c$ are single, and OTUs with $\theta > \theta_c$ are composed.

Method for identifying OTUs with noise amongst single ones

We focus now on OTUs identified as single. Later, we will discuss possible tools to split OTUs identified as being composed in order to obtain a clustering of the sample's reads formed only of single OTUs. In order to determine if a single OTU contains noise reads or not, we propose a fully automatic supervised classification method whose input variables are features derived from the dissimilarity array D_{otu}. Namely, we use a linear Support Vector Machine (SVM) to discriminate between the two types of single OTUs. To derive the features, we estimate the parameters of a Stochastic Block Model (SBM, Holland et al. (1983); Daudin et al. (2008); Lee and Wilkinson (2019)) with a Poisson distribution on dissimilarities and with two blocks (see Suppl. material 1: A for a description of the SBM model). The reason for choosing two blocks is that, in the presence of noise, we expect that the core reads of the OTU will be grouped into one block and the atypical sequences into another. We use the block parameters as features. More precisely, the SBM makes it possible to cluster individuals, based on their pairwise dissimilarities. Individuals in the same block share the same pattern of connectivity. A specificity of SBM (as opposed to classical clustering methods) lies in its plasticity: a block is not necessarily assortative (i.e. with small within-block dissimilarities); it may also be dissortative. Our argument for choosing SBM is that, if there are some noise reads in an OTU, they will be grouped into a dissortative block. If there is no noise, the two blocks will be assortative. These two different patterns can be identified using the connectivity matrix Λ of the SBM. In the case of a two-block SBM, this is a 2 x 2 symmetric matrix. The two diagonal elements, Λ(1, 1) and Λ(2, 2), correspond to the mean intra-block dissimilarity and the non-diagonal element Λ(1, 2) corresponds to the mean inter-block dissimilarity. If a single OTU contains noise sequences, they will be grouped into one of the two blocks, let us say, block 2, with a large value for Λ(2, 2) and for Λ(1, 2). If the OTU is without noise, all the diagonal elements of Λ should be small. We chose the two values, Λ(1, 2) and *max*(Λ(1, 1), Λ(2, 2)), as features for the linear SVM. We considered other combinations of the elements of Λ, but they did not improve the performance of the classification and this choice is easier to interpret in terms of presence/absence of noise.

In practice, we assigned an ’expert’ label to each OTU of a training set, amongst ’with noise’, ’uncertain’ and ’without noise’. To do this, we computed the normalised degree β_{seq} of each read of the OTU, defined as the percentage of dissimilarities smaller than the barcoding gap in the row corresponding to this read in the dissimilarity array D_{otu}: $\beta_{seq} = 100\, \Sigma_{i \neq j}\, \delta\, (d\,(i, j) \leq g)/(n_{otu} - 1)$. If the minimum of β_{seq} over the OTU reads is lower than 20%, the OTU is labelled as ’with noise’; if it is larger than 70%, it is labelled as ’without noise’; otherwise, the OTU is labelled as ’uncertain’. Only OTUs labelled as ’with noise’ or ’without noise’ are used to learn the SVM. Note that this method could be directly envisaged as a candidate for identifying OTUs with noise. However, it is not fully automatic since it relies on two thresholds that were manually defined and some OTUs remain unclassified (’uncertain’ type). We refer to it as the degree-based classifier below.

Overview of the whole procedure

We summarise here the succession of steps to perform when using our method to identify composed, single with noise and single without noise OTUs of a set of samples. We define two sets of samples:

- a set S of samples, where we want to type each OTU of each sample as composed, single with noise, single without noise;
- for doing that, another set T of samples, different from S, used as a training set to learn the SVM classifier.

The identification as composed/simple of each OTU is an unsupervised classification which is done sample per sample. The identification as with/without noise for a single OTU is done OTU by OTU with a SVM classifier which is learned on the training set T. Knowing that, here are the steps for typing all samples in S:

1. Learning an SVM classifier with a linear kernel for typing as with/without noise all single OTUs in all samples in T, by:
    i. deriving the subset of all single OTUs in T, denoted T_{single}, using the histogram of the θs,
    ii. assigning an expert label (with/without noise) to each OTU, in T_{single} using the degree-based classifier on G_{otu} (the graph of the OTU where two reads are linked by an edge if their dissimilarity is smaller than the barcoding gap),
    iii. computing the connectivity matrix Λ_{otu} of each OTU as a 2-class clustering of each distance array D_{otu} with an SBM, assuming a Poisson distribution of the distances,
    iv. learning an SVM classifier with linear kernel with two features $\Lambda(1, 2)$ and $max(\Lambda(1, 1), \Lambda(2, 2))$ of the Λ matrices of all OTUs in T.
2. typing all OTUs in all samples s in S:
    i. first as composed/single, using the histogram of the θs in s,
    ii. second, for all single OTUs, as with/without noise using the SVM classifier built in step 1 and the SBM clustering of each OTU.

All the steps in the above procedure are elementary and can be written with any language (like python or R). We provide in a Figshare project and a gitlab project (see Section Data Accessibility) a set of programmes which assemble them in a given way and which we used for producing our results. Other solutions are possible and equivalent. The gitlab provides a documentation of the programmes and a tutorial on the assemblage we propose on a subset of the complete dataset (all samples) to save time and memory while running it. It gives some guidelines for the user who wishes to use the programmes on his/her own datasets.

Results

Identification of composed OTUs in diatom samples

An expert classification can be built, based on visual inspection, in order to validate the output of our identification method. However, we did not built it on the whole dataset since it would require a visual inspection of 2529 dissimilarity arrays. We built it only for one location. We chose the Teychan location, since this location will also be used as a training set for the identification of OTUs with noise in a second step. The Teychan dataset, therefore, refers to the set of the eight samples located at Teychan (two samples per season: one for benthic and one for pelagic). It is composed of 654 OTUs.

Here, we first present the validation of the method for identifying composed OTUs, on the samples located at Teychan. Then, on the whole dataset, we tested the existence of a link between OTU type and OTU size and we analysed the assignation pattern of composed OTUs.

Analysis of the Teychan dataset

For the Teychan dataset, an expert classification of each OTU of each sample into one of the three categories - composed, single or uncertain – has been built, based on an expert procedure which works as follows. First, heatmaps of the dissimilarity array of each OTU were drawn, with reads ordered according to the leaves of a dendrogram (Aggregative Hierarchical Clustering with Ward criteria, Müllner (2013)). Examples of a heatmap for one entity only and for two entities are given in Fig. 1. Second, the graph G attached to a composed OTU is organised as a set of connected communities, one per entity. Such graphs were drawn for each OTU of the sample. Finally, we attributed the character “single” or “composed” for each OTU by visual inspection of the heatmap (presence or absence of a block structure) and the graph (presence or absence of communities). Most of the cases were unambiguous, clearly belonging to one type or the other. However, the transition is smooth rather than discrete (for example, θ varies continuously). It may happen that intermediate cases occur, for example, if there are two entities with highly unbalanced sizes, like a dominant one and a small satellite one. In such a case, the OTU was labelled as “uncertain”.

The contingency table built from the 654 OTUs of the Teychan dataset (Table 1) shows a very good agreement between the expert classification and the automatic one. In particular, single OTUs are very well identified: only 12 false negatives (amongst 104 composed OTUs) and only nine false positives

Table 1. Comparison of the expert classification and the automatic classification of the OTUs into the composed and single categories, for the Teychan dataset.

		Automatic		
		Composed OTUs	**Single OTUs**	**Total**
Expert	Composed OTUs	92	12	104
	Uncertain OTUs	11	12	23
	Single OTUs	9	518	527
Total		112	542	654

Table 2. Link between OTU size and its classification as composed or single. Statistics of the ranks (the ranks are ordered from smallest to largest size) and the p-value of the Wilcoxon Mann-Whitner test.

Number of OTUs	**2529**
Mean rank for single OTUs	1163.5
Mean rank for composed OTUs	1778.5
p-values	1.535×10^{-55}

(amongst 527 single OTUs). Uncertain OTUs are evenly distributed between composed and single categories by the automatic method. In Suppl. material 1: D, we provide the individual contingency matrices and θ_c values for each of the eight samples composing the Teychan dataset.

Link between OTU type and OTU size

We then applied the procedure to the whole dataset (the 32 samples). We tested the hypothesis of a link between the OTU type (single or composed) and its size. Suppl. material 1: C, fig. S2 visually presents the link between the size of an OTU and its type. It can be seen that single OTUs (green and blue dots) have small to medium sizes and that composed OTUs (red dots) have larger sizes. This was quantified by a Wilcoxon Mann-Whitney test (function mannwhitneyu() in Python library scipy.stats) between single and composed categories. The results (see Table 2) show strong evidence for a link between the OTU size and its type (composed or single).

Link with assignation

Amongst the 180 OTUs that were fully annotated with a taxon (see Section Data), eight were categorised as composed. We observed three situations. For two of them, there are two or three species present in the OTU and the dissimilarity array D_{otu} and graph G_{otu} are clearly structured into two blocks separating one species from the other(s). This is the typical situation that we target when identifying composed OTUs. Three other OTUs are monospecific and there is no obvious structure in D_{otu} or G_{otu}. However, they have the particularity that reads are loosely connected to the others, leading to a large value of θ, larger than θ_c. Finally, the last three OTUs are monospecific (or nearly) and D_{otu} and G_{otu} are nevertheless structured into two blocks. An example of each situation is given in Suppl. material 1: C, fig. S1.

Identification of OTUs with noise in diatom samples

Training on the Teychan dataset

The method to identify OTUs with noise is a supervised method that requires a training set to learn the SVM. The most discriminant factors when studying community diversity are the season and the water column. This is the reason why we built the training set on one location (Teychan) and the test set on the other three locations. Both sets contain samples associated with different and balanced values for the season and the water column. This training step is performed using only OTUs that have been categorised as with or without noise by the degree-based classifier (uncertain OTUs cannot be used here).

For each choice of features (pair of coefficients of the Λ matrix), we ran a 10-fold cross validation to estimate the error of prediction. We obtained the best Area Under Curve value (AUC = 0.951) with the features $f_1 = max(\Lambda(1, 1), \Lambda(2, 2))$ and $f_2 = \Lambda(1, 2)$. The feature f_1 represents the mean dissimilarity between two reads of the SBM block with the larger mean intra dissimilarity. If there are noise reads, they should be in this block. The feature f_2 represents the mean inter-block dissimilarity in the SBM model. The SVM classifier frontier is defined by the expression $y = 9.452 + 0.569\, f_1 + 0.876\, f_2$. Contingency Table 3 reports the comparison between the two classification methods, now including the OTUs categorised as uncertain by expertise (the eight contingency matrices, one per sample in the Teychan dataset, are provided in Suppl. material 1: D). The SVM classifier very efficiently detects the OTUs with noise (only six missed amongst 381). It is a bit less efficient to detect OTUs without noise (six missed amongst 48). The majority of uncertain OTUs are classified as being with noise by the SVM classifier.

Results on the test set

The SVM classifier obtained on the training set is applied to the OTUs of the 24 samples of the test set (i.e. all samples, except those in the Teychan dataset). Since the expert method can also be automated, we can compare the results of the two classifiers. They are reported in contingency Table 4. For both methods, there are much fewer OTUs classified as without noise than with noise. The SVM classifier identifies all the OTUs with noise. However, only 64% of the OTUs without noise are identified. We can see (Fig. 4) a pattern in the values of the two features, f_1 and f_2 that are used by the SVM classifier, depending on the OTU type (with or without noise). We recall that f_1 represents the mean intra dissimilarity of the SBM block with larger intra dissimilarity. f_2 is the mean inter block dissimilarity in the inferred SBM model. Single OTUs identified as 'without

Table 3. Comparison of the degree-based classification and the SVM classification of the single OTUs into the 'with noise' and 'without noise' categories, on the Teychan dataset (training set).

		Automatic		
		OTUs with noise	OTUs without noise	Total
Expert	OTUs with noise	375	6	381
	Uncertain OTUs	87	26	113
	OTUs without noise	6	42	48
Total		468	74	542

Table 4. Comparison of the degree-based classification and the SVM classification of the single OTUs into the 'with noise' and 'without noise' categories, on the test set.

		Automatic		
		OTUs with noise	**OTUs without noise**	**Total**
Expert	OTUs with noise	1228	0	1228
	Uncertain OTUs	277	24	301
	OTUs without noise	16	29	45
Total		1521	53	1574

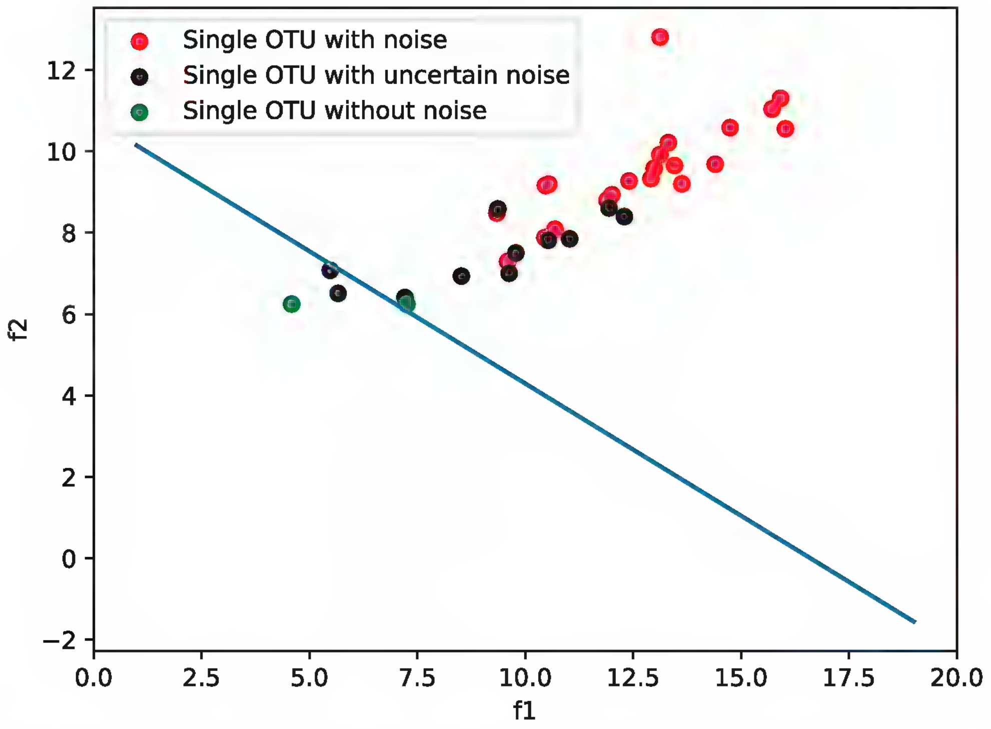


Figure 4. SVM frontier for the Comprian-Pelagic-Autumn sample. The two axes represent the values of the two features used for classification. A dot represents an OTU and is coloured according to the expert classification.

noise' are associated with a low value of f_1 (between 4 and 8) and a low value of f_2 (between 4 and 8 as well). On the contrary, single OTUs identified as 'with noise' are associated with large values of f_1 (almost always between 6 and 16) and with large values of f_2 (between 6 and 25). Furthermore, these two parameters of the inferred SBM model increase simultaneously, showing a gradient of noise intensity amongst the single OTUs.

Link between OTU size and OTU type

For OTUs categorised as single, we test the hypothesis of a link between the OTU size and its category (with or without noise). The Wilcoxon Mann-Whitney test has been used (based on the single OTUs of the 32 samples) and the results show that there is strong evidence for such a link (see Table 5).

Table 5. Link between OTU size and its classification as single with or without noise. Statistics of the ranks (the ranks are ordered from smallest to largest size) and the p-value of the Wilcoxon Mann-Whitney test.

Number of OTUs	**2116**
Mean rank for single OTUs without noise	552.5
Mean rank for single OTUs with noise	1089.7
p-values	3.7x 10^{-22}

Link with assignation

Amongst the 180 OTUs that were fully annotated, 153 were categorised as single with noise and 23 as single without noise. Ignoring the artefactual presence of sequences of *Rhizosolenia fallax* species, almost all were monospecific (only two exceptions).

Link between sample composition and environmental conditions

Having applied the two procedures to each of the 32 samples balanced for season, location, water column for identification of composed, single with noise and single without noise OTUs, we computed the proportion of each type per sample. In Fig. 5, we show a visualisation by ternary plot of these proportions. Globally, all samples have a low proportion of OTUs without noise and we can observe that the proportion of single OTUs with noise and composed OTUs vary from one sample to another.

The central ternary plot of Fig. 5 suggests a potential link between these proportions and the water column of the sample. This is also suggested by the plot of the fraction of composed OTUs for each of the 32 samples as displayed in Suppl. material 1: C, fig. S3. To test this link, we first considered two sets of 16 values: the list of percentages of composed OTUs in the benthic samples and in the pelagic samples. We applied a Wilcoxon rank test and obtained a p-value of 1.6 x 10^{-4}. The mean fraction of composed OTUs in a benthic sample (respectively a pelagic sample) is 0.19 (respectively 0.10). Consequently, there is strong evidence that the fraction of composed OTUs is larger in benthic samples than in pelagic ones.

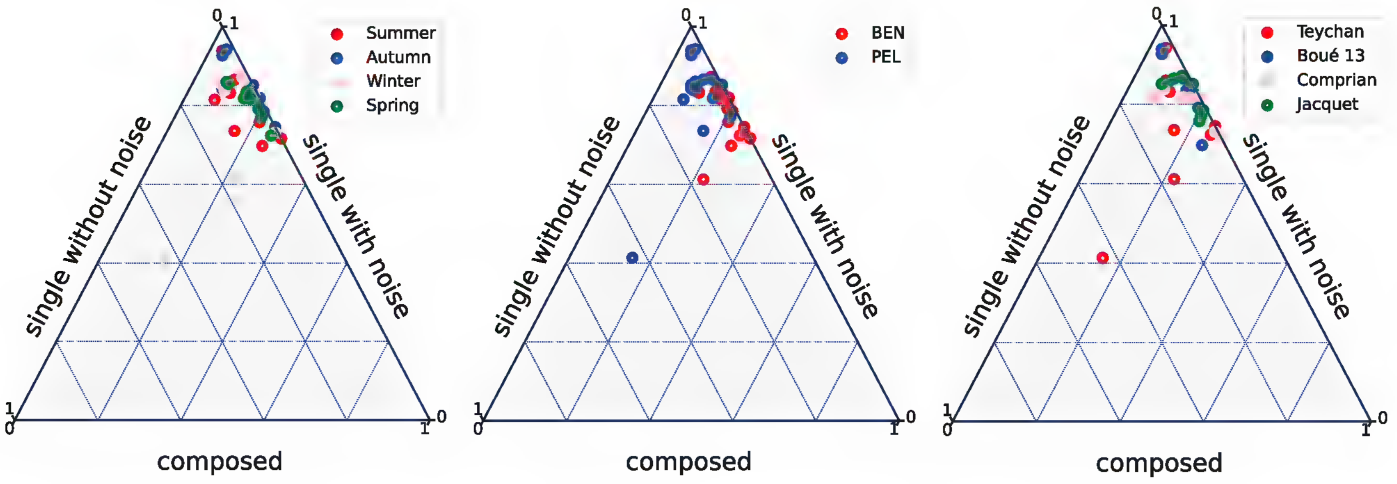


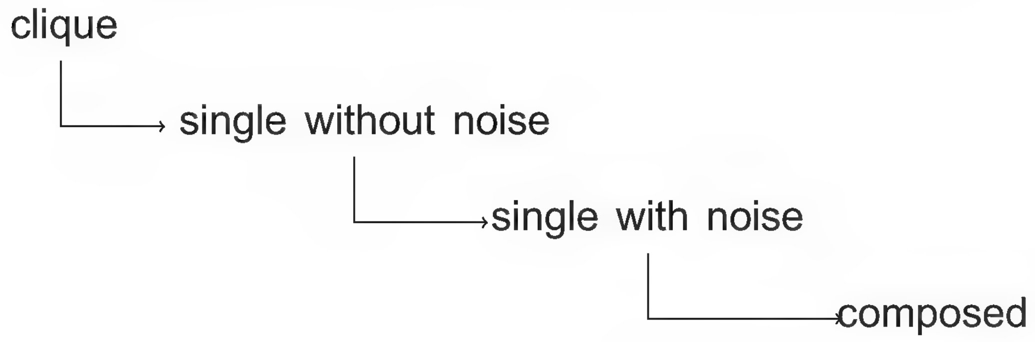
Figure 5. Visualisation of the proportion of composed, single with noise and single without noise OTUs for each sample. Left: dots coloured by seasons, centre: dots coloured by water column, right: dots coloured by location.

We then considered two other sets of 16 values: the list of percentage of OTUs with noise (amongst the single OTUs) in the benthic samples and in the pelagic samples. We also applied a Wilcoxon rank test and we obtained a p-value of 0.04763. We concluded that there is no evidence that the fraction of OTUs with noise (amongst the single OTUs) in a sample is different for benthic and pelagic conditions.

We did not test whether the other environmental conditions (season, location) have or do not have an influence on the composition in the sample since the number of observations per condition would be too small (8).

Discussion

The discussion of the quality of the different types of OTUs is organised along a gradient of complexity of the structure of the OTUs, as follows:



Cliques

The expected structure of the graph G_{otu} built from the dissimilarity array D_{otu} is a clique if the dissimilarities are the age of the Most Recent Common Ancestor (MRCA). In such an ideal case, the OTU is obviously reliable. However, in practice, we work with evolutionary distances computed from local alignment scores. The discrepancy between the age of the MRCA and evolutionary distances within a set of sequences increases with the age of the MRCA. It can, therefore, be expected that cliques represent clusters with a relatively young MRCA and that the evolutionary distances within the cluster are closely related to the age of the MRCA. This allows us to postulate that cliques built from evolutionary distances are OTUs of good quality. There are four cliques over all of the 32 samples of diatoms. Three of them have no annotated reads. This may mean that they represent species that are absent from the reference database. One of them is partially annotated, always with the same species. The fact that some reads in the clique are not recovered probably means that mapping reached its limit in terms of quality, because if a query maps on references with different taxa, the mapping is said to be ambiguous and the read is not annotated.

Single OTUs without noise

Let us recall that the noise (or the absence of noise) in a single OTU is detected, based on the value of two features f_1 and f_2, where f_1 represents the mean dissimilarity between two sequences of the SBM block with the largest mean intra-block dissimilarity and f_2 represents the mean inter-block dissimilarity. A single OTU is typed as “without noise” if the parameters f_1 and f_2 are both small, as illustrated in Fig. 4.We showed that both features f_1 and f_2 are always lower than 8 for single OTUs without noise. This implies that, for those OTUs in the SBM modelling of D_{otu}, all dissimilarities are realisations of a Poisson distribution with a mean lower than the barcoding gap (nine in our study). Therefore, the graph G_{otu} is close to a clique. It is tempting to extrapolate and derive the conclusion that single OTUs without noise can be considered of good quality for use in further studies.

In order to provide in what follows indications about the quality of an OTU (which means it can be accepted as an OTU for further studies) that is not a clique nor a single OTU without noise, we referred to an external expert evaluation. Although we are agnostic as to whether an OTU has or does not have a taxonomic meaning, we used the mapping of reads on a reference database as external information. If all the reads in an OTU are annotated and assigned to the same species, then OTU picking and taxonomy converge, suggesting that the OTU can be considered of good quality. Otherwise it is questionable. Hence, we focused on fully annotated OTUs in the rest of the discussion.

Single OTUs with noise

A single OTU is typed as "with noise" if features f_1 and f_2 are both large. Such an OTU displays a minority of satellite reads, which are close to (at a distance smaller than the gap) only a small fraction of the remaining reads (the core, the main densely connected entity). In the subsample of fully annotated OTUs, almost all of the single OTUs with noise are monospecific ones, regardless of the quantity or intensity of noise. Whether such a conclusion can be extended beyond fully annotated OTUs is an open question and deserves further studies on a diversity of organisms to progress along this line. Indeed, a partial covering only by mapping can be due to the fact that uncovered reads either belong to another species absent in the reference database, lowering the acceptability of the OTU or that they belong to the same species, but are labelled as unknown due to imperfections and errors in the mapping or the reference database.

Composed OTUs

Composed OTUs are very likely to be large OTUs and to be composed of two or more entities each of which is a candidate to be a more reliable OTU. However, in the subsample of fully annotated OTUs, we observed some composed OTUs with a different profile: either monospecific OTUs with, overall, a low level of connections in G_{otu} or monospecific OTUs with a clear structure divided into two blocks. Both cases lead to large values of missing edges and the OTUs are, therefore, typed as composed. In the latter case, one possible reason for the block pattern of the dissimilarity array may be a structure in the intraspecific molecular diversity. However, the number of specimens in one OTU is often too small to check with population genetics indices (see Phillips et al. (2018) for a discussion about the sample size necessary for assessing molecular intraspecific diversity). Regarding large composed OTUs, the production of spurious clusters by a chaining effect in aggregative clustering with single linkage is well known (see, for example, Kopp (1978)) and can lead to composed OTUs. Programmes like SWARM (Mahé et al. 2014, 2015) have identified this issue and provide a way to solve it by breaking the chains in the amplicon space. Here, we suggest that chaining can occur because of the non-universality of the barcoding gap. Some structures of the dissimilarity array of a sample are more likely to lead to chaining. These situations raise the question the separability of OTUs. We suggest that this is an issue for the quality of OTUs. Some examples of the diversity of those situations are given in Suppl. material 1: C, fig. S4. Such a variability can be understood keeping in mind that the barcoding gap is not universal. We refer the reader to Phillips et al. (2022), the running title of which is "Is the Barcoding Gap Real?" for a thorough discussion and critique of the notion of barcoding gap. It can vary between clades in a sample. Indeed, let us assume that, for a given small gap, we have a set of well-delineated OTUs corresponding each to a taxon. Then increasing the gap to build connected components will lead to new edges between former OTUs and, possibly, to a network of connected entities, with a weakening of the possibility to discriminate them, as in the top graph of Suppl. material 1: C, fig. S4. We have shown that composed OTUs are the largest ones with very high significance. This means that medium size and small size OTUs are single and likely to correspond to taxa. This also means that the selected barcoding gap is relevant for delineating most OTUs (the middle and small size ones), but inadequate for most of the largest ones: the existence of spurious composed OTUs reflects the inadequacy of the selected barcoding gap to delineate OTUs amongst those sequences. This may explain the difference in the ratio of single/composed OTUs between benthic (with several composed large OTUs) and pelagic samples (with fewer ones): the structure of the molecular diversity of pelagic and benthic diatom flora differ. Indeed, most species are preferentially present either in the benthic samples or in the pelagic samples (see Suppl. material 1: C, fig. S5). We can hypothesise from this observation a pattern where distances between species in benthic flora are globally smaller than in pelagic flora. Being benthic or pelagic probably affects the rhythms of alternation between sexual and asexual modes of reproduction and may, therefore, have an impact on patterns of DNA molecular diversity. This influence can only be indirect, given that we have worked with rbcL, which is a chloroplastic marker. This deserves further investigation.

Finally, the large spurious OTUs, automatically detected by $\theta > \theta_c$, should be reshaped as sets of new and smaller OTUs. Two ways to do this are to build them as outcomes of either unsupervised clustering of the dissimilarity array of the composed OTU or of community detection (see Fortunato (2010) for a review of this approach) in the graph induced by the array.

Conclusion

Recent advances in massively parallel sequencing technology has led to the rapid production of millions of reads. This has opened the way to the analysis of many environmental communities, leading to further exploration of their diversity and ecology, at a pace that was unimaginable beforehand. The building blocks of such studies are sets of OTUs obtained by clustering the reads of a given sample. In this context of massive data, it is no longer possible to scrutinise each OTU one by one to assess its quality and decide to keep it or not, or to reshape it. Here, we propose a tool to make progress in assessing automatically the quality of an OTU, with OTUs streaming through a pipeline. It relies on the comparison between the OTU's inner structure (given as its pairwise dissimilarity array) and an ideal one and by characterising two ways in which the structure of an OTU can deviate from the ideal situation: first, we distinguish composed vs. single OTUs. Second, amongst the single OTUs, we distinguish OTUs with and without noise. We applied the method on 32 samples of diatoms collected in Arcachon Bay (France) that represent contrasted environmental conditions and we obtained good agreement with expert categorisation of OTUs. We suggest that single OTUs without noise can be used as such for further ecological studies. Composed OTUs should be post-treated with classical clustering of community detection tools. The quality of single OTUs with noise remains to be further tested via supplementary studies on a diversity of organisms.

Our method can be implemented in a pipeline and used automatically and sequentially on a large number of OTUs belonging to one or different samples.

This builds a quality filter that enhances the reliability of subsequent studies in ecology and diversity structures that are undertaken on these same data, by strengthening their foundations.

Furthermore, the impact of the dissimilarities and classification methods on the OTUs quality deserves further investigation and the optimal choice can depend on the sample studied. Our tool could also provide a way to identify, for a given sample, the dissimilarities and classifications methods that lead to the set of OTUs with the best intrinsic quality, for example, distances computed from alignment scores (see Gusfield (1997)).

Acknowledgements

We thank the participants of the Malabar project for their authorisation to use the data produced in this project, especially the Laboratoire Environnement-Ressources of IFREMER at Arcachon for the field campaign, Emilie Chancerel and Franck Salin at INRAE BioGeCo for the production of DNA sequences. Computer time for the preparation of the data in this study was provided by the computing facilities of the MCIA (Mésocentre de Calcul Intensif Aquitain). The dissimilarity arrays have been produced in the Malabar project, supported by ”Cote Labex” Call for Research Projects, Year 2017.

Additional information

Conflict of interest

The authors have declared that no competing interests exist.

Ethical statement

No ethical statement was reported.

Funding

No funding was reported.

Author contributions

A. Franc: conceptualisation, methodology, validation, formal analysis, investigation, original draft writing, review and editing. N. Peyrard: conceptualisation, methodology, validation, formal analysis, investigation, original draft writing, review and editing. M.-J. Cros: software, validation, formal analysis, investigation, review and editing. J.-M. Frigerio: data curation, review and editing.

Author ORCIDs

Marie-Josée Cros https://orcid.org/0000-0002-6395-5563
Jean-Marc Frigerio https://orcid.org/0000-0003-0471-2075
Nathalie Peyrard https://orcid.org/0000-0002-0356-1255
Alain Franc https://orcid.org/0000-0001-9448-8569

Data availability

The codes for learning the noise classifier and for determining the type of OTUs are available in the GitLab project https://forgemia.inra.fr/alain.franc/otu_shape, where the user can find a documentation and a tutorial on a smaller dataset than the one used in our study. The code and the data to replicate our study are available in a Figshare project (Cros et al. 2022), with reference link https://doi.org/10.6084/m9.figshare.20764690.v3. The dissimilarity arrays are publicly available as well at https://doi.org/10.57745/7T2UCB (Malabar project).

References

Auby I, Méteigner C, Rumebe M, Chancerel E, Salin F, Aluome C, Barraquand F, Carassou L, Del Amo Y, Meleder V, Petit A, Picoche C, Frigerio JM, Franc A (2022) Malabar datasets used in study “OTU quality from dissimilarity arrays”. Recherche Data Gouv, V1. https://doi.org/10.57745/7T2UCB

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. Trends in Ecology & Evolution 27(4): 233–243. https://doi.org/10.1016/j.tree.2011.11.010

Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 360(1462): 1935–1943. https://doi.org/10.1098/rstb.2005.1725

Cortes C, Vapnik V (1995) Support-vector networks. Machine Learning 20(3): 273–297. https://doi.org/10.1007/BF00994018

Cox T, Cox MAA (2001) Multidimensional Scaling. In: Chapman Hall/CRC (Eds) Monographs on Statistics and Applied Probability, 2nd edn., Vol. 88, 328 pp. https://doi.org/10.1201/9780367801700

Cros MJ, Frigerio JM, Peyrard N, Franc A (2022) Code, dataset and results for the study “OTU quality from dissimilarity arrays”. Figshare. https://doi.org/10.6084/m9.figshare.20764690.v3

Daudin JJ, Picard F, Robin S (2008) A mixture model for random graphs. Statistics and Computing 18(2): 173–183. https://doi.org/10.1007/s11222-007-9046-7

Fortunato S (2010) Community detection in graphs. Physics Reports 486(3-5): 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Frigerio JM, Rimet F, Bouchez A, Chancerel E, Chaumeil P, Salin F, Thérond S, Kahlert M, Franc A (2016) Diagno-syst: a tool for accurate inventories in metabarcoding. arXiv. https://arxiv.org/abs/1611.09410

Froslev T, Kjoller R, Bruun H, Ejrnaes R, Brunbjerg A, Pietroni C, Hansen A (2017) Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. Nature Communications 8(1): 1188. https://doi.org/10.1038/s41467-017-01312-x

Girvan M, Newman M (2002) Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America 99(12): 7821–7826. https://doi.org/10.1073/pnas.122653799

Gower JC, Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis. Applied Statistics 18(1): 54–64. https://doi.org/10.2307/2346439

Gusfield D (1997) Algorithms on Strings, Trees and Sequences. Cambridge University Press, 534 pp. https://doi.org/10.1017/CBO9780511574931

Hajibabaei M, Shokralla S, Zhou X, Singer GAC, Baird DJ (2011) Environmental barcoding: A next generation sequencing approach for biomonitoring applications using river benthos. PLOS ONE 6(4): e17497. https://doi.org/10.1371/journal.pone.0017497

Holland P, Laskey K, Leinhardt S (1983) Stochastic blockmodels: First steps. Social Networks 5(2): 109–137. https://doi.org/10.1016/0378-8733(83)90021-7

Kermarrec L, Franc A, Rimet F, Chaumeil P, Humbert JF, Bouchez A (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: A test for freshwater diatoms. Molecular Ecology Resources 13(4): 607–619. https://doi.org/10.1111/1755-0998.12105

Kopp B (1978) Hierarchical Classification I. Biometrical Journal. Biometrische Zeitschrift 20(5): 495–501. https://doi.org/10.1002/bimj.4710200506

Lee C, Wilkinson D (2019) A review of stochastic block models and extensions for graph clustering. Applied Network Science 4: 122. https://doi.org/10.1007/s41109-019-0232-2

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: Robust and fast clustering method for amplicon-based studies. PeerJ 2: e593. https://doi.org/10.7717/peerj.593

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M (2015) Swarm v2: Highly-scalable and high-resolution amplicon clustering. PeerJ 3: e1420. https://doi.org/10.7717/peerj.1420

Müllner D (2013) fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python. Journal of Statistical Software 53(9): 1–18. https://doi.org/10.18637/jss.v053.i09

Phillips JD, Gillis DJ, Hanner RH (2018) Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. Ecology and Evolution 9(5): 2996–3010. https://doi.org/10.1002/ece3.4757

Phillips JD, Gillis DJ, Hanner RH (2022) Lack of statistical rigor in DNA barcoding likely invalidates the presence of a true species' barcode gap. Frontiers in Ecology and Evolution 10: 859099. https://doi.org/10.3389/fevo.2022.859099

Rimet F, Chaumeil P, Keck F, Kermarrec L, Vasselon V, Kahlert M, Franc A, Bouchez A (2016) R-Syst:diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. Database (Oxford) 2016: baw016. https://doi.org/10.1093/database/baw016

Taberlet P, Coissac E, Hajibabaei M, Rieseberg L (2012) Environmental DNA. Molecular Ecology 2(8): 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Zinger L, Lionnet C, Benoiston AS, Donald J, Mercier C, Boyer F (2021) metabaR: An R package for the evaluation and improvement of DNA metabarcoding data quality. Methods in Ecology and Evolution 12(4): 586–592. https://doi.org/10.1111/2041-210X.13552

Supplementary material 1

Supplementary information

Authors: Marie-Josée Cros, Jean-Marc Frigerio, Nathalie Peyrard, Alain Franc

Data type: pdf

Explanation note: **A** stochastic Block Model **B** estimation of θ density for composed OTU identification **C** figures **D** tables.

Link: https://doi.org/10.3897/mbmg.8.108649.suppl1